Received: 9 November 2022 | Revised: 4 April 2023 | Accepted: 29 May 2023

DOI: 10.1002/sim.9824

RESEARCH ARTICLE

Statistics in Medicine WILEY

Deep causal feature extraction and inference with neuroimaging genetic data

Yuchen Yao[1] | **Dipnil Charkraborty**[2] | **Lin Zhang**[2] | **Xiaotong Shen**[1] | **Alzheimer's Disease Neuroimaging Initiative** | **Wei Pan**[2]

[1]School of Statistics, University of Minnesota, Minneapolis, Minnesota, USA

[2]Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA

Correspondence
Wei Pan, Division of Biostatistics, University of Minnesota, A460 Mayo Building, MMC 303, Minneapolis, Minnesota, USA.
Email: panxx014@umn.edu

Funding information
National Institutes of Health, Grant/Award Numbers: R01AG065636, R01AG069895, U01AG073079; the Minnesota Supercomputing Institute

Alzheimer's disease (AD) is a severe public health issue in the world. Magnetic Resonance Imaging (MRI) offers a way to study brain differences between AD patients and healthy individuals through feature extraction and comparison. However, in most previous works, the extracted features were not aimed to be causal, hindering biological understanding and interpretation. In order to extract causal features, we propose using instrumental variable (IV) regression with genetic variants as IVs. Specifically, we propose Deep Feature Extraction via Instrumental Variable Regression (DeepFEIVR), which uses a nonlinear neural network to extract causal features from three-dimensional neuroimages to predict an outcome (eg, AD status in our application) while maintaining a linear relationship between the extracted features and IVs. DeepFEIVR not only can handle high dimensional individual-level data for model building, but also is applicable to GWAS summary data to test associations of the extracted features with the outcome in subsequent analysis. In addition, we propose an extension of DeepFEIVR, called DeepFEIVR-CA, for covariate adjustment (CA). We apply DeepFEIVR and DeepFEIVR-CA to the Alzheimer's Disease Neuroimaging Initiative (ADNI) individual-level data as training data for model building, then apply to the UK Biobank neuroimaging and the International Genomics of Alzheimer's Project (IGAP) AD GWAS summary data, showcasing how the extracted causal features are related to AD and various brain endophenotypes.

KEYWORDS

Alzheimer's disease, causal inference, instrumental variable, MRI, neural networks

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1 | INTRODUCTION

Alzheimer's disease (AD) leads to memory loss, cognitive dementia and behavioral changes, which caused over 121 thousand deaths in 2019.[1,2] In 2022, it is estimated that there are 6.5 million AD patients among the American elderly aging 65 or above and this number is predicted to be 13.8 million by 2060.[1] Due to the prevalence and impact of AD, an increasing number of researchers work on AD data collection and analysis. The Alzheimer's Disease Neuroimaging Initiative (ADNI)[3] dataset is one of the most comprehensive and widely used datasets for AD. Structural magnetic resonance imaging (MRI) scans are the major part of this dataset, which are three-dimensional images showing brain structures and region conditions. In addition to MRI scans, the ADNI dataset also contains genome-wide Single-Nucleotide Polymorphisms (SNPs) for each individual, which are genetic variants nowadays commonly used in genome-wide association studies (GWAS).

Under the reasonable assumption that the pathway of AD is from genetics (and environmental factors) to brain atrophy, then to AD, it has been advocated to use neuroimaging features as endophenotypes to gain statistical power because of their closer proxy to causal genetic factors in the AD pathway, motivating the development of the field of neuroimaging genetics. In addition to boosting power, the use of endophenotypes may provide important clues about causal pathways to the disease. A recent GWAS demonstrated the effectiveness of this strategy: some risk genes such as *FRMD6*, were first identified to be associated with some neuroimaging intermediate phenotypes, for example, hippocampal atrophy,[4] then were later validated to be associated with AD.[5,6] More generally, there may be other regions of interest (ROIs) or even other imaging features as more effective endophenotypes waiting to be discovered.

Existing neuroimaging GWAS are almost all based on *manually* extracted or *predefined* imaging features as endophenotypes, for example, based on some ROIs from a brain atlas.[7–11] However, due to limited knowledge, there is still debate on how to define ROIs or even brain atlases; furthermore, these ROIs may or may not be most relevant for the given GWAS trait, that is, AD here.[12] As a motivating question, there are 66 existing atlases for the whole brain structural MRI data:[12] which one to use? It would be of high interest to develop and apply data-driven methods for novel feature extraction, especially given the recent success of deep learning in image analysis. Due to high dimensionality of MRI scans, feature extraction is both necessary and challenging. A tensor regression method and a scalable algorithm are proposed to reduce dimension from high dimensional predictors.[13] Tensor Partition Regression Models (TPRM) are designed to combine extracted features from partitioned tensors.[14] In recent years, deep learning techniques have been utilized in feature extraction. A convolutional auto-encoder network is proposed to reduce dimension from MRI scans.[15,16] A recent study proposes to extract features by training 3D convolutional neural networks in an AD classification task, reporting an AUC score of 0.75 on test data.[17] They also perform GWAS scans to identify genetic variants associated with extracted features. However, due to the existence of *hidden* confounders that affect both brain images and the AD status (or other outcomes), the above methods are not designed for and cannot be interpreted as causal feature extraction, which may hinder biological understanding of AD mechanisms.

Instrumental variable (IV) regression is a popular tool to learn causal relationships between some exposures and an outcome while being robust to hidden confounding. In the ADNI dataset, the exposures are MRI scans or their extracted features, the outcome is the AD status, and the SNPs can be treated as IVs. Two-stage least squares (2SLS) is the most widely used method in IV regression.[18] 2SLS estimates the exposures using IVs in the first stage, and uses the estimated exposures to predict the outcome in the second stage. Linear regression models are fitted in both stages. When genetic variants/SNPs are used as IVs, some special cases of IV regression (notably with independent SNPs) are often known as Mendelian randomisation (MR) that have recently been applied to neuroimaging data.[19,20] One of the most popular MR is inverse-variance weighted Mendelian randomisation (IVW-MR),[21] designed for one exposure and combining results from each SNP by inverse-variance weighting. Multi-variable Mendelian randomisation (MVMR) is an extension to multiple exposures.[22] To lift the restriction of the linearity assumption, nonparametric models like kernel methods and basis functions are used in IV regression.[23,24] A nonparametric method is proposed to detect nonlinear causal effects of a scalar/univariate exposure on the outcome based on GWAS summary data in transcriptome-wide association studies (TWAS).[25] Compared to classical nonparametric models, neural networks are an alternative choice for flexible modeling of nonlinear relationships. DeLIVR keeps a linear regression model in the first stage but fits a neural network in the second stage.[26] DFIV[27] and DeepIV[28] apply neural networks in both stages to learn nonlinear causal relationships. DeepGMM[29] is based on the generalized method of moments (GMM) using nonlinear functions of exposures and IVs.

Although the existing IV regression methods can extract causal features, they are not ideal in some applications as for the ADNI dataset. First, for methods fitting a linear model in the first stage like 2SLS and DeLIVR, modeling a 3D image by a linear function is not effective in general. Besides, we use MRI scans in the ADNI-1 dataset which is a part of the ADNI dataset and contains MRI scans of 817 individuals and even fewer if the genetic data are combined. Thus, we want to use all samples in the ADNI-1 dataset for training and validation and use AD GWAS summary statistics for hypothesis testing. AD GWAS summary statistics are based on (marginal) linear models between AD and SNPs. In general, nonlinear regression models that do not guarantee a linearity between IVs and the outcome cannot be applied to GWAS summary data. There are some similarities between our models/assumptions and those in a previous study,[25] however, a key difference is that we apply CNNs to 3-dimensional images as a high-dimensional exposure while they deal with a scalar (ie, gene expression) exposure.

In order to apply IV regression to high dimensional exposures and use GWAS summary statistics in testing at the same time, we propose a novel method called Deep Feature Extraction via Instrumental Variable Regression (DeepFEIVR). For features extracted by a convolutional neural network from high dimensional exposures, DeepFEIVR projects them onto the space of IVs and uses these projected features to predict the outcome, thus ensuring both the relevance/predictivity and causal interpretation of the extracted features for the outcome. In Section 2, we first introduce 2SLS, then discuss the details of our proposed method DeepFEIVR and its extension for covariate adjustment, DeepFEIVR-CA. Section 3 shows some simulation results of DeepFEIVR. In Section 4, we extract some causal features from MRI scans in the ADNI dataset by DeepFEIVR and test for possible associations between the extracted features and AD using a large-scale AD GWAS summary dataset, followed by a comparison with the results of DeepFEIVR-CA. Then we explore the relationships between the extracted features and brain regions/endophenotypes before ending with a short discussion.

2 | METHODS

2.1 | Notation

Assume that $Z \in \mathbb{R}^p$, $X \in \mathbb{R}^k$, and $Y \in \mathbb{R}$ represent IVs, exposures and the outcome respectively. X and Y are affected by a hidden (combined) confounder U. In the IV regression setting, we consider a training set $\mathcal{D}_{tr} = \{\mathbf{Z}_{tr}, \mathbf{X}_{tr}, \mathbf{Y}_{tr}\}$ with size n and a validation set $\mathcal{D}_{val} = \{\mathbf{Z}_{val}, \mathbf{X}_{val}, \mathbf{Y}_{val}\}$ with size m. In hypothesis testing, we first consider an individual-level test set $\mathcal{D}_{te}^{in} = \{\mathbf{Z}_{te}, \mathbf{Y}_{te}\}$ with size n'. Based on this individual-level test set, we can create a set of summary statistics $\mathcal{D}_{te}^{s} = \{(\hat{\gamma}_j, \widehat{\mathrm{Var}}(\hat{\gamma}_j)) : j = 1, \dots, p\}$. γ_j is the effect size of the jth element of Z on Y, which can be estimated by a linear model regressing $\mathbf{Y}_{te}$ on the jth column of $\mathbf{Z}_{te}$. The summary statistics also include $\widehat{\mathrm{Var}}(\hat{\gamma}_j)$ for the jth IV, which is the squared standard error of γ_j. Under some weak assumptions, the summary statistics can be used in hypothesis testing in place of the individual-level test set.

2.2 | An existing method: 2SLS

We start with introducing 2SLS. The causal model structure of 2SLS is

$$\begin{aligned} \text{stage 1}: \quad & X = B^{\intercal} Z + U_1 + \Delta_1, \\ \text{stage 2}: \quad & Y = \beta^{\intercal} X + U_2 + \Delta_2, \end{aligned}$$

where $U_1 \in \mathbb{R}^k$ and $U_2 \in \mathbb{R}$ are correlated confounders, $\Delta_1 \in \mathbb{R}^k$ and $\Delta_2 \in \mathbb{R}$ are two error terms with zero means and constant variances, and Z is independent of (U_1, Δ_1). For notational convenience, we leave out the intercepts in this paper and assume Z, X, Y are already centered at sample mean 0. $B \in \mathbb{R}^{p \times k}$ and $\beta \in \mathbb{R}^k$ are the parameters to be estimated. It is assumed that there exist confounders affecting both X and Y, so U_1 and U_2 are correlated while Δ_1 and Δ_2 are independent. In order to eliminate the impact from confounders, Z (IVs) should satisfy the following three assumptions: (1) the distribution of the exposure X given IVs Z is not constant in Z; (2) Z is independent of the outcome Y conditional on X, U_2 and Δ_2; (3) Z and (U_2, Δ_2) are independent. In the first stage of 2SLS, we estimate the mean of X by a linear model regressing X on Z. In the second stage, we estimate Y by a linear model regressing Y on $\hat{X}$, where $\hat{X}$ is the estimated mean of X obtained from the first stage. In hypothesis testing, the null hypothesis is $\beta_j = 0$ for some or all of $j = 1, 2, \dots, k$, where β_j is a component of $\beta = (\beta_1, \beta_2, \dots, \beta_k)$. Since Y and Z are assumed to follow a linear relationship in 2SLS, we can use summary statistics in hypothesis testing. However, 2SLS is not effective in modeling high dimensional exposures by using a linear model in the first stage, so in the next part, we consider a deep learning based instrumental variable regression method.

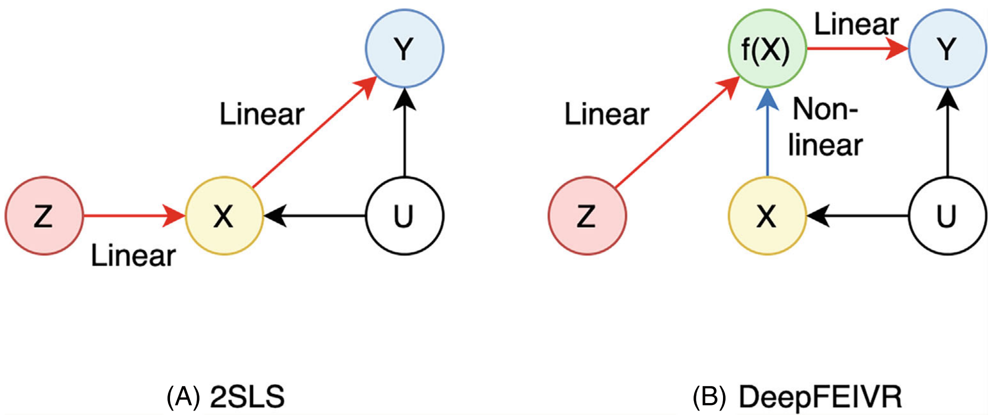

FIGURE 1 Causal model comparison between 2SLS and DeepFEIVR.

2.3 | New method: Deep feature extraction via instrumental variable regression

2.3.1 | Causal model structure

The causal model structure of Deep Feature Extraction via Instrumental Variable Regression (DeepFEIVR) involves two stages

$$\begin{aligned} \text{stage 1}: \quad & f(X) = B^\intercal Z + U_1 + \Delta_1, \\ \text{stage 2}: \quad & Y = \beta^\intercal f(X) + U_2 + \Delta_2, \end{aligned} \tag{1}$$

where f is a nonlinear multivariate function extracting q features from X. $B \in \mathbb{R}^{p\times q}$, $\beta \in \mathbb{R}^q$, and f are estimated in the first and second stages. We assume $q < p$. $U_1 \in \mathbb{R}^q$ and $U_2 \in \mathbb{R}$ are correlated. $\Delta_1 \in \mathbb{R}^q$ and $\Delta_2 \in \mathbb{R}$ are defined as before in 2SLS: Δ_1 and Δ_2 are independent and they both have zero means and constant variances. Z and (U_1, Δ_1) are independent. For simple notation Z and Y are assumed to be centered at mean 0. In DeepFEIVR, three IV assumptions are revised by replacing the exposure X with the features $f(X)$: (1) the conditional distribution of the features $f(X)$ given Z is not constant in Z; (2) Z is independent of Y conditioning on $f(X)$, U_2 and Δ_2; (3) Z and (U_2, Δ_2) are independent. The first assumption says that Z is relevant to $f(X)$, and the second assumption can be directly satisfied if Z, X and Y follow the causal model structure (1). Taking expectation of Y conditional on X gives $\mathbb{E}(Y|X) = \beta^\intercal f(X) + \mathbb{E}(U_2 + \Delta_2|X)$ where $\mathbb{E}(U_2 + \Delta_2|X) \neq 0$. Neglecting this nonzero term can lead to incorrect inference. Thus, IVs are used for correct inference about β.

In the first stage, $\mathbb{E}(f(X)|Z) = B^\intercal Z$, so we can project $f(X)$ onto the space of Z and estimate $\mathbb{E}(f(X)|Z)$ by $\hat{B}^\intercal Z$. Then in the second stage, $\mathbb{E}(Y|Z) = \mathbb{E}(\beta^\intercal f(X)|Z) = \beta^\intercal \mathbb{E}(f(X)|Z)$ and we can estimate the conditional expectation of Y given Z by replacing $\mathbb{E}(f(X)|Z)$ with $\hat{B}^\intercal Z$. With B and β being estimated, $\mathbb{E}(Y|Z)$ is estimated by $\hat{\beta}^\intercal \hat{B}^\intercal Z$, which is a linear transformation of Z. Thus, summary statistics can be used in subsequent hypothesis testing in DeepFEIVR as to be discussed in Section 2.3.3.

2SLS can be regarded as a special case of DeepFEIVR by replacing $f(X)$ with X in causal model (1). We compare the causal models of 2SLS and DeepFEIVR in Figure 1.

2.3.2 | Estimation

In order to capture the nonlinearity of f, we use $f_\theta \in \mathcal{F} = \{f_\theta : \theta \in \Theta\}$ to model f, where $\mathcal{F}$ is a class of neural networks with a specified architecture, and estimate β and θ in batches in a stochastic gradient descent (SGD)-type algorithm. With a batch set $\{\mathbf{Z}_b, \mathbf{X}_b, \mathbf{Y}_b\}$ of size n_b, we update the estimates of β and θ by

$$\min_{\theta,\beta} \frac{1}{n_b} \|\mathbf{Y}_b - \mathbf{Z}_b \hat{B}_\theta^b \beta\|_2^2 + \Omega(\theta, \beta), \tag{2}$$

where $\hat{B}_\theta^b = \left(\mathbf{Z}_b^\intercal \mathbf{Z}_b + \lambda n_b I\right)^{-1} \mathbf{Z}_b^\intercal f_\theta(\mathbf{X}_b)$ is a closed-form solution to the minimization problem in the first stage:

$$\min_B \frac{1}{n_b} \|f_\theta(\mathbf{X}_b) - \mathbf{Z}_b B\|_2^2 + \lambda \|B\|_2^2.$$

This is a ridge regression model with a ridge penalty parameter λ while the minimization problem (2) is for the second stage. $\Omega(\theta, \beta)$ is an elastic net regularization term for θ and β.

2.3.3 | Hypothesis testing

After training the model, we obtain the estimated θ and then the estimated weight matrix $\hat{B}_{\hat{\theta}} = \left(\mathbf{Z}_{tr}^\intercal \mathbf{Z}_{tr} + \lambda n I\right)^{-1} \mathbf{Z}_{tr}^\intercal f_{\hat{\theta}}(\mathbf{X}_{tr})$. Note that, this estimate is based on the entire training set. For a global testing of whether there is any association between the features and the outcome, we test

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0 \;\text{ versus }\; \mathrm{H}_1 : \text{there exists } j \text{ such that } \beta_j \neq 0.$$

For an individual test on the jth feature, we test

$$\mathrm{H}_0 : \beta_j = 0 \;\text{ versus }\; \mathrm{H}_1 : \beta_j \neq 0.$$

Next, we show how to perform hypothesis testing based on $\hat{B}_{\hat{\theta}}$ with an individual-level test dataset $\mathcal{D}_{te}^{in}$ and a set of summary statistics $\mathcal{D}_{te}^{s}$.

Given an individual-level test set $\mathcal{D}_{te}^{in} = \{Z_{te}, Y_{te}\}$, it is straightforward to perform a global Wald test on the association between Y_{te} and $Z_{te}\hat{B}_{\hat{\theta}}$, and individual Wald tests on that between Y_{te} and each column of $Z_{te}\hat{B}_{\hat{\theta}}$, to detect (putative) causal relationships between the features and the outcome. When the size of the individual-level test set is not large enough, Wald tests can be inaccurate.

In case that large individual-level data are not available, we can consider using GWAS summary statistics calculated based on some large but unavailable individual-level dataset. Assume that a set of summary statistics $\mathcal{D}_{te}^{s} = \{\hat{\gamma}_j, \widehat{\mathrm{Var}}(\hat{\gamma}_j)\}_{j=1}^p$ with the sample size n' is available. As in previous studies,[31] we can estimate the coefficient vector β_S and its covariance matrix using the following formulas

$$\hat{\beta}_S = \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1} \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y, \tag{3}$$

$$\widehat{\mathrm{Var}}(\hat{\beta}_S) = \frac{1}{n' - q} \left(Y^\intercal Y - \hat{\beta}_S^\intercal \hat{B}_{\hat{\theta}}^\intercal Z^\intercal Y\right) \left(\hat{B}_{\hat{\theta}}^\intercal Z^\intercal Z \hat{B}_{\hat{\theta}}\right)^{-1}. \tag{4}$$

We can estimate $Z^\intercal Z$ by $\frac{n'}{n_R} Z_R^\intercal Z_R$ where $Z_R \in \mathbb{R}^{n_R \times p}$ is a matrix of Z in a reference panel. This reference dataset can be the ADNI dataset or a separate independent individual-level genotype dataset. Note that $Z^\intercal Y \in \mathbb{R}^p$ and denote the jth element of $Z^\intercal Y$ as $\{Z^\intercal Y\}_j$. Its estimate $\widehat{\{Z^\intercal Y\}}_j$ can be taken as $\widehat{\{Z^\intercal Z\}}_{jj} \hat{\gamma}_j$, where $\widehat{\{Z^\intercal Z\}}_{jj}$ is the jth diagonal element of the estimated $\{Z^\intercal Z\}$. The median of the set $\left\{(n'-1)\widehat{\{Z^\intercal Z\}}_{jj} * \widehat{\mathrm{Var}}(\hat{\gamma}_j) + \hat{\gamma}_j \widehat{\{Z^\intercal Y\}}_j, j = 1, 2, \dots, p\right\}$ can be used to estimate $\{Y^\intercal Y\}$. Based on $\hat{\beta}_S$ and $\widehat{\mathrm{Var}}(\hat{\beta}_S)$, a Wald test is used to test q features globally. For $j = 1, 2, \dots, q$, based on $\hat{\beta}_{S,j}$ and $\widehat{\mathrm{Var}}(\hat{\beta}_S)_{jj}$, one can test whether the jth extracted feature is associated with Y. Although the above formulas for $\hat{\beta}_S$ and $\widehat{\mathrm{Var}}(\hat{\beta}_S)$ are derived for a quantitative Y, as discussed in previous studies,[31] we can still use these formulas for a binary Y when using GWAS summary data.

The extracted causal features are not unique. For any invertible matrix $M \in \mathbb{R}^{q \times q}$, we can rewrite the two-stage models as: $M^{-1} f(X) = M^{-1} B^\intercal Z + M^{-1} U_1 + M^{-1} \Delta_1$ and $Y = \beta^\intercal M M^{-1} f(X) + U_2 + \Delta_2$. Thus, $M^{-1} f(X)$ equivalently represents the causal features with the corresponding association parameter $\beta^\intercal M$. However, this does not affect testing whether the extracted features are associated with the outcome because $\beta = 0$ is equivalent to $M^\intercal \beta = 0$.

2.4 | New method: DeepFEIVR-CA

We extend the proposed method by covariate adjustment (CA) when some covariates are present in the training and validation sets. We consider the following causal model

$$\begin{aligned} \text{stage 1}: &\quad f(X) = B^{\intercal}Z + A^{\intercal}W + U_1 + \Delta_1, \\ \text{stage 2}: &\quad Y = \beta^{\intercal}f(X) + \gamma^{\intercal}W + U_2 + \Delta_2, \end{aligned} \tag{5}$$

where $W \in \mathbb{R}^w$ are covariates, $A \in \mathbb{R}^{w\times q}$ and $\gamma \in \mathbb{R}^w$ are the unknown parameters for covariates. We assume (Z, W) are independent of (U_1, Δ_1) and three IV assumptions become: (1) the conditional distribution of $f(X)$ given Z and W is not constant in Z and W; (2) Z is independent of Y conditional on $f(X)$, W, U_2 and Δ_2; (3) (Z, W) and (U_2, Δ_2) are independent. In the first stage, we have $\mathbb{E}(f(X)|Z, W) = B^{\intercal}Z + A^{\intercal}W$, and in the second stage, $\mathbb{E}(Y|Z, W) = \beta^{\intercal}\mathbb{E}(f(X)|Z, W) + \gamma^{\intercal}W$. Based on (5), we propose DeepFEIVR-CA for covariate adjustment. For parameter estimation in DeepFEIVR-CA, we modify (2) as

$$\min_{\theta,\beta,\gamma} \frac{1}{n_b}\|\mathbf{Y}_b - (\mathbf{Z}_b, \mathbf{W}_b)\hat{B}^b_{\theta}\beta - \mathbf{W}_b\gamma\|_2^2 + \Omega(\theta, \beta), \tag{6}$$

where $\mathbf{W}_b$ is the matrix for covariates in a batch set and $\hat{B}^b_{\theta} = ((\mathbf{Z}_b, \mathbf{W}_b)^{\intercal}(\mathbf{Z}_b, \mathbf{W}_b) + \lambda n_b I)^{-1}(\mathbf{Z}_b, \mathbf{W}_b)^{\intercal}f_{\theta}(\mathbf{X}_b)$. Then $\hat{B}_{\hat{\theta}}$ is estimated as $((\mathbf{Z}_{tr}, \mathbf{W}_{tr})^{\intercal}(\mathbf{Z}_{tr}, \mathbf{W}_{tr}) + \lambda nI)^{-1}(\mathbf{Z}_{tr}, \mathbf{W}_{tr})^{\intercal}f_{\theta}(\mathbf{X}_{tr})$.

With individual-level test data, hypothesis testing is straightforward. However, GWAS summary data usually do not offer any information about covariates; assuming that Z and W are nearly uncorrelated, we can conduct hypothesis testing as in DeepFEIVR without covariates.

2.5 | CNN model architecture

In DeepFEIVR as applied to the simulated data and the ADNI data, f is a nonlinear function estimated by a CNN. For real data, the input of this CNN is 3D MRI images with 3 channels for white matter, gray matter and cerebrospinal fluid (WM, GM, CSF) respectively, and the extracted features are the output of this CNN model. The architectures of DeepFEIVR and the direct CNN applied in simulations and real data (ADNI) are shown in Figures 2 and 3. More details and model architecture of DeepFEIVR-CA are provided in the Appendix.

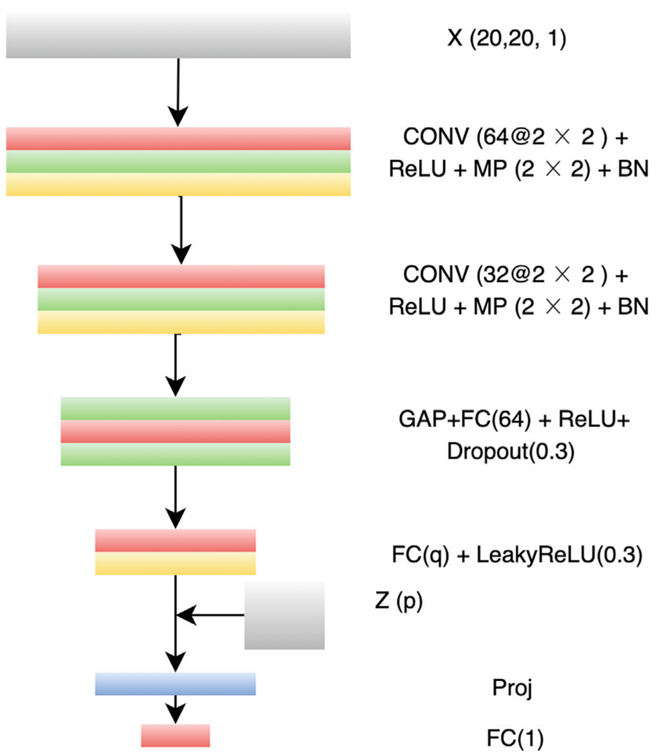

FIGURE 2 The model architecture of DeepFEIVR applied to the simulation dataset. Z are IVs and X are simulated images.

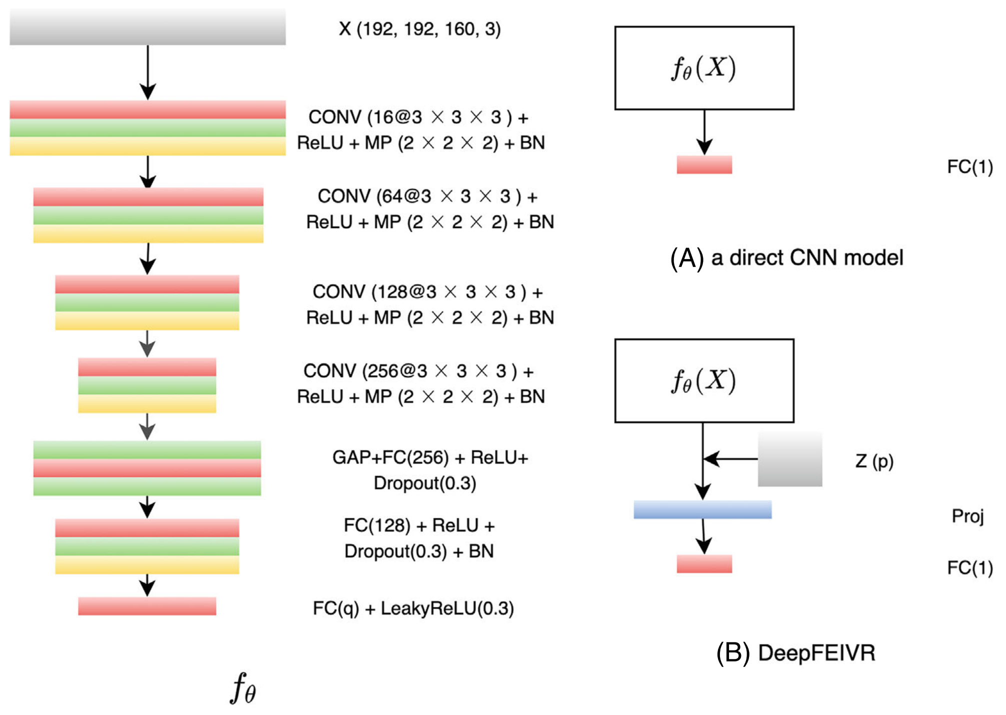

FIGURE 3 The model architecture of DeepFEIVR used in the ADNI dataset. Left: the model architecture f_θ. Top-Right: the direct CNN model in which a linear regression model follows f_θ. Bottom-Right: DeepFEIVR in which the extracted features by f_θ are projected onto the column space of the IVs and then a linear regression model is applied. Z and X are the IVs and input image respectively.

For a neural network trained on images, gradCAM[31] is a technique to detect important regions of images from the network perspective. In a CNN neural network including a Global Averaging Pooling (GAP) layer, suppose $\{Z_{whd} \in \mathbb{R}^{p_1}, w = 1, 2, \dots, n_w, h = 1, 2, \dots, n_h, d = 1, 2, \dots, n_d\}$ are outputs of any layer before the GAP layer. Then the activation of voxel (w, h, d) for the jth feature is calculated by $|A^j_{whd}|$ with $A^j_{whd} = Z^\top_{whd}\left(\frac{1}{n_w n_h n_d}\sum_{whd}\frac{\partial F_j}{\partial Z_{whd}}\right)$, where F_j is the jth extracted feature. The original version of gradCAM is designed for a classification problem and uses $\mathrm{ReLU}(A^j_{whd})$ as the activation function. Since the extracted features are not binary here, we use the absolute function instead.

3 | SIMULATIONS

In this part, simulations are conducted to assess the performance of DeepFEIVR. In each replicate of simulation, we first simulate $Z \in \mathbb{R}^{50}$, partitioned into 10 groups with 5 IVs in each group. Each group of IVs is generated from a multivariate normal distribution with correlation 0.1 between different IVs. The IVs between different groups are independent. U and ϵ_2 are generated from $N(0, 1)$ independently, while ϵ_{11} and ϵ_{12} are generated from $N(0, 0.25)$ independently. Then X, f and Y are generated by

$$
\begin{aligned}
f &:= \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = B^\top Z + \begin{pmatrix} U \\ U \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \end{pmatrix}, \\
X &= \mathrm{Image}(f_1, f_2) + N, \\
Y &= \beta^\top f + U + \epsilon_2,
\end{aligned}
$$

where each element in $B \in \mathbb{R}^{50\times 2}$ is generated by a standard normal distribution independently, and $\beta = d * (-0.1, 0.2)^\top$ with $d \in \{0.00, 0.02, 0.03, 0.05\}$. $X \in \mathbb{R}^{20\times 20}$ is an image generated from f by presenting two squares in the left and right parts with values $\sqrt{|f_1|}$ and $\sqrt{|f_2|}$ respectively; the signs of f_1 and f_2 determine that the squares appear in the top or bottom parts of the image. If f_1 is positive, the square for f_1 appears in the left-top part of the image; otherwise, the square appears

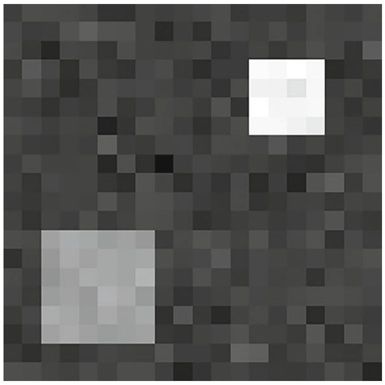

FIGURE 4 An image example (gray scale) with $f_1 = -1$ and $f_2 = 4$. The square for f_1 (with size 6×6) is positioned at the left-bottom of the image as f_1 is negative while the square for f_2 (with size 4×4) is at the right-top part as f_2 is positive.

in the left-bottom part. The square for f_2 is similarly determined (in the right part of the image). The sizes of two squares are independently sampled from $\{4 \times 4, 6 \times 6, 8 \times 8\}$. N is a 20×20 matrix, each element of which follows $N(0, 0.01)$ independently. See Figure 4 for an example. Based on how X is generated, in return f can be viewed as a function of X.

We repeat each simulation setup for 100 times, and in each replicate we apply DeepFEIVR using individual–level data or summary statistics, 2SLS using individual–level data or summary statistics and a direct CNN. In the direct CNN, we use the output of the second last layer in a CNN as the extracted features and use them to perform hypothesis testing. In 2SLS, we regress vec(X) (ie, the vector of all elements of X) on Z in the first stage and conduct hypothesis testing of Y on $\hat{B}^{\top}_{\mathrm{vec}(X)}Z$ (as the estimate mean of vec(X)). In the second stage, due to high dimension of vec(X), the degrees freedom of the Wald test is estimated by the effective rank of $\hat{B}^{\top}_{Vec(X)}Z$. For DeepFEIVR, we do *not* assume that we know the number of the true features (ie, 2 in the simulated data). So in the standard setting, we choose 4 as the (estimated) number of extracted features (q) and use the batch size 32. In each replicate, training data size is 800, validation data size is 200, test data size is 4000 (the same data set used for hypothesis testing with individual-level data or summary statistics) and the size for reference panel is 20000.

We show the results of the sample proportions of the p-values smaller than 0.05 from 100 replicates for DeepFEIVR, the direct CNN and 2SLS in Table 1. In addition to the standard setting, we also consider the situations of silencing a feature in simulations (with $\beta = d * (0.0, 0.2)^{\top}$), using weak IVs (with the elements in the first 25 rows of B generated from $N(0, 0.01)$ independently), using the estimated number of features as $q = 2$ (same as in the simulated data), using a smaller batch size of 16, or using a doubled training sample size of 1600.

As shown in Table 1, for DeepFEIVR, the estimated Type I error rates ($d = 0.0$) are close to 0.05, while the estimated power ($d > 0$) grows with increasing d. The results based on the summary statistics are close to those on the individual-level data, confirming that it is good to use summary statistics for hypothesis testing in DeepFEIVR. In contrast, the direct CNN fails to extract causal features, which is expected as it does not use IVs to distinguish true causal features from hidden confounding effects. Furthermore, DeepFEIVR outperforms 2SLS regardless of $\beta = d * (-0.1, 0.2)^{\top}$ or $\beta = d * (0.0, 0.2)^{\top}$, or of the presence of weak IVs. For DeepFEIVR, using a smaller batch size or using a different number of features does not make much difference, and its power grows slightly with the larger training sample size.

4 | REAL DATA ANALYSES

4.1 | Datasets

4.1.1 | ADNI

In the real data analysis, the SNPs, MRI scans and labels of AD status in both the training and validation data as well as a brain region of interest (ROI) dataset are downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (https://adni.loni.usc.edu). In 2003, ADNI, guided by Michael W. Weiner, MD, was started and involved collaboration from private and public sources. The original label in the ADNI dataset indicates CN (normal control), MCI (mild cognitive impairment) and AD (Alzheimer's disease). The principal objective of ADNI was to study the

TABLE 1 Simulation results of DeepFEIVR, the direct CNN and 2SLS in simulations: the empirical type I errors (for $d = 0.0$) and power (for $d > 0$) at the nominal significance level of 0.05 based on 100 independent replicates for each setup.

	Test data	Methods	$d = 0.0$	$d = 0.02$	$d = 0.03$	$d = 0.05$
Standard	Individual	DeepFEIVR	0.05	0.11	0.33	0.74
	Summary	DeepFEIVR	0.05	0.10	0.33	0.74
	Individual	Direct CNN	0.97	1.00	1.00	1.00
	Individual	2SLS	0.08	0.11	0.08	0.28
	Summary	2SLS	0.07	0.10	0.06	0.30
$\beta = d * (0.0, 0.2)^{\intercal}$	Individual	DeepFEIVR	0.04	0.15	0.32	0.65
	Summary	DeepFEIVR	0.04	0.14	0.33	0.64
	Individual	2SLS	0.01	0.08	0.18	0.24
	Summary	2SLS	0.02	0.07	0.19	0.24
Weak IVs	Individual	DeepFEIVR	0.08	0.08	0.13	0.39
	Summary	DeepFEIVR	0.08	0.07	0.15	0.40
	Individual	2SLS	0.09	0.06	0.09	0.15
	Summary	2SLS	0.10	0.07	0.09	0.22
$q = 2$	Individual	DeepFEIVR	0.05	0.09	0.27	0.70
	Summary	DeepFEIVR	0.05	0.09	0.28	0.70
Batch size 16	Individual	DeepFEIVR	0.06	0.13	0.30	0.80
	Summary	DeepFEIVR	0.06	0.12	0.29	0.80
Training size 1600	Individual	DeepFEIVR	0.05	0.13	0.38	0.77
	Summary	DeepFEIVR	0.07	0.12	0.39	0.77

progression of AD by combining MRI scans, genetic data, as well as other neuropsychological test results. Updated information can be found at https://www.adni-info.org. In this paper, we combine MCI and CN and set the binary label to be AD or not AD. 755 individuals in the ADNI dataset have both MRI scans and SNP data available. In the 755 individuals, 175 individuals are labeled as AD. Before applying DeepFEIVR, we reduce the number of IVs (SNPs) to 317 by the following steps:

1. Remove SNPs that not appeared in the reference panel (UK Biobank individual-level genotypic data) and two summary statistics sets IGAP and BIG40. The details of these datasets will be introduced later.
2. SNPs with a missing value proportion smaller than 20% are selected for further consideration and for each selected SNP, we impute NAs with the mode.
3. We fit a linear model regressing the binary label (AD or not AD) on each SNP and remove SNPs with a p-value larger than a cutoff 0.001.
4. For a pair of highly correlated SNPs with the absolute correlation larger than 0.8, the SNP with a larger p-value is removed.

After GradWarp correction, B1 nonuniformity correction and N3 bias field correction, ADNI provides 1075 T1-weighted 1.5 T 3D structural MRI scans*. In preprocessing of MRI scans, we first extract brain tissues using Brain Extraction Tool (BET)[32] and then use FMRIB's (Functional Magnetic Resonance Imaging of the Brain) Automated Segmentation Tool (FAST)[33] to partition brain tissues into three types, white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). Both tools are from FSL (FMRIB Software Library). To use a common dimension of the images from different individuals while preserving much information for learning, we crop all 3D images into the dimension/shape of (192, 192, 130) and normalize voxel values inside [0, 1]. We treat the three tissue types as 3 channels and use the CNN architecture in Section 2.5 to extract features from the combined 3D images with 3 channels. Figure 5 shows an

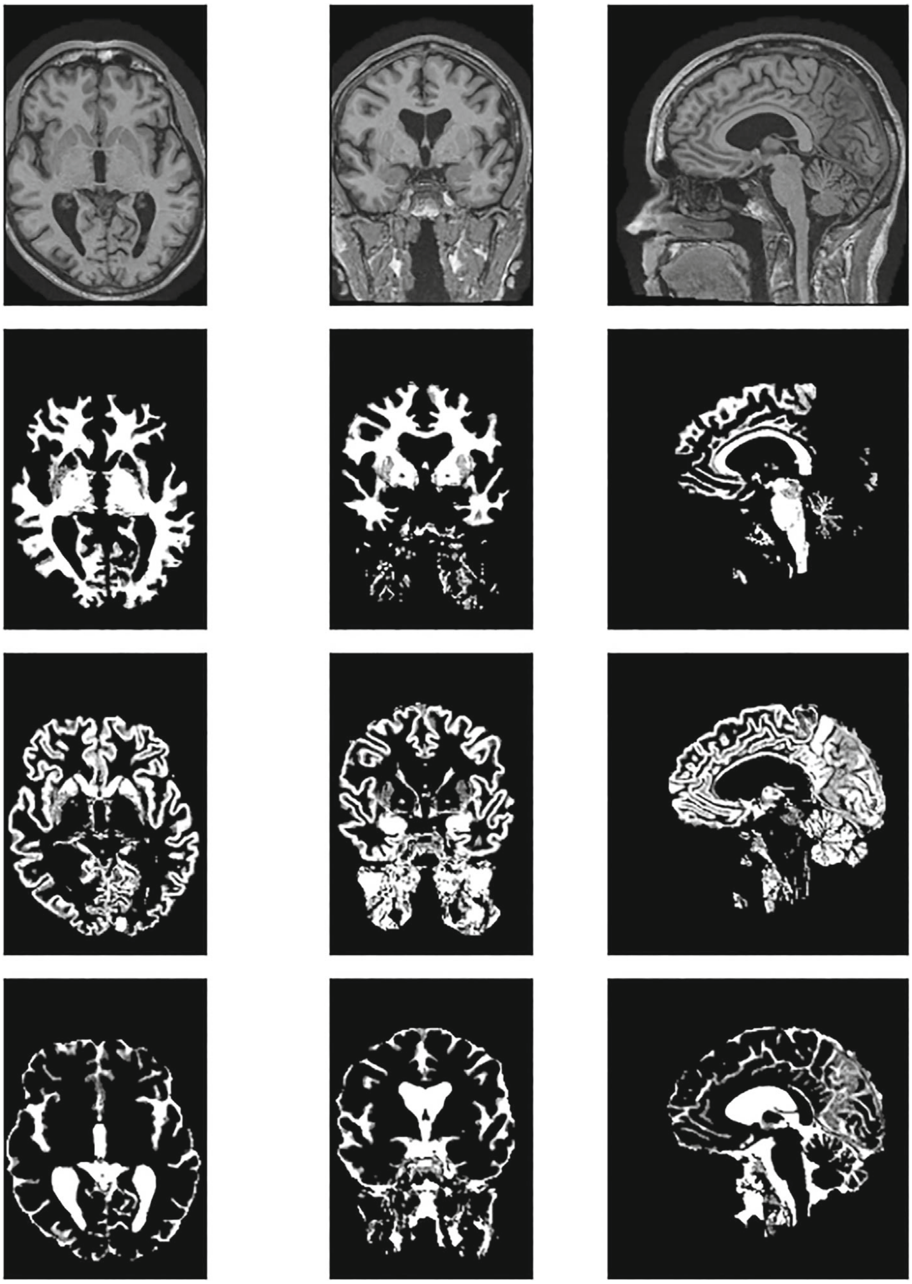

FIGURE 5 An example of the original MRI scan, WM, GM and CSF tissues (from the top row to the bottom) in coronal (left), axial (middle) and sagittal (right) planes in the ADNI dataset. The repetition time is 2400 ms and echo time is 3.5 ms. The flip angle is 8 degree.

image sample and its 3 tissues from 3 planes. In addition to MRI and genetic data, the ADNI dataset also provides a Region of Interest (ROI) dataset, which records gray matter volumes of 116 brain regions from gray matter maps of the ADNI images generated by longitudinal Voxel-Based Morphometry (VBM). The regions are determined using Anatomical Automatic Labelling (AAL).[34] We can use this dataset to study how our extracted causal features are related to brain regions.

In the ADNI data, 3D MRI scans are treated as the high dimensional exposure and a binary label (AD or not) is the outcome. Directly extracting features by fitting a neural network model on MRI scans to predict AD cannot guarantee that the extracted features are causal. For example, age can be one of the confounders that change conditions in brain regions and directly affect the risk of AD at the same time. The changes in brain regions can be highly correlated with AD but we cannot claim that these changes are all causal contributors. We treat some SNPs as IVs since SNPs are likely to satisfy the three valid IV assumptions: some SNPs are associated to the brain, may influence the risk of AD only through the brain, and are independent of hidden confounders. Finally, some baseline variables, such as age, gender and handedness, can be used as (observed) covariates/confounders.

4.1.2 | IGAP

The limited sample size of the ADNI dataset drives us to find another dataset for association testing. Due to the privacy issues in genetic datasets, large datasets containing individual-level genetic and AD status information are not easily accessible. Fortunately, several AD GWAS summary datasets based on large AD GWAS studies are publicly available. The International Genomics of Alzheimer's Project (IGAP)[35] provides a summary statistics dataset we need for hypothesis testing. IGAP conducts two stages of AD GWAS studies and they provide summary statistics of two stages separately. We use the summary statistics from Stage 1 as it involves more individuals and more SNPs. In Stage 1, the summary statistics are calculated based on 17 008 AD individuals and 37 154 control individuals for over 7 million SNPs. To avoid individual identification, IGAP only provides the estimated effect sizes of SNPs on AD and their corresponding standard errors.

4.1.3 | UK Biobank individual-level genotypic data

UK Biobank individual-level data[37] contains genetic and other information on around 490 000 individuals of age 40+ from 2006 to 2010. We can use this dataset as the reference panel for genotypes in hypothesis testing with GWAS summary data; that is, we use it to obtain a robust estimate of $Z^{\intercal}Z$.

4.1.4 | BIG40

The Oxford Brain Imaging GWAS Data (BIG40), as a part of the UK Biobank data, collects the summary statistics of 3935 imaging-derived phenotypes (IDPs) in a GWAS study performed on around 33k individuals, a subset of the UK Biobank participants. For 3935 IDPs and over 17 million SNPs, the (marginal) effect size of each SNP on each IDP and its standard error are estimated. The IDPs are numeric measurements derived from multi-modal MRI scans (eg, the volume of a specific brain region). We will use this set of summary statistics to identify significant IDPs related to the extracted features by DeepFEIVR.

4.2 | Extracted features and their association with AD

In training the direct CNN model and the DeepFEIVR model, the batch size is chosen to be 16 and 672 samples are in the training set. We use the rest samples as validation samples. In hypothesis testing of DeepFEIVR, we use the IGAP AD GWAS summary statistics dataset as summary data, and the UK Biobank individual-level genotypic data as the reference panel. A Wald test is conducted on 20 extracted features. The p-value of this Wald test is 8.321×10^{-11}, indicating that the extracted 20 features together are significantly associated with AD. Next we conduct individual tests on each of the 20 features separately. Negative $\log_{10}$ (p-values) of the 20 individual tests are shown in Figure 6, in which the 4th, 5th, 6th, 11th, 13th, 17th, 19th, and 20th features are significant for AD at a p-value threshold 0.05.

As in previous neuroimaging GWAS studies with ADNI data,[7] for DeepFEIVR-CA, we consider the baseline age, gender and handedness as covariates; each of the covariates is not significantly associated with the IVs as shown in a linear model regressing each covariate on all IVs. The p-value of the global test for AD association with the extracted features based on the IGAP AD GWAS summary statistics is 0.039, which is statistically significant but less significant than that from DeepFEIVR; we will comment more on this result in the Discussion section. To assess how the extracted features by DeepFEIVR and DeepFEIVR-CA are related, we conduct a canonical correlation analysis on the two sets of the extracted features. Canonical correlations measure some maximum similarities between two sets of the features through their linear combinations. The ith canonical correlation is the maximum correlation between a linear combination of causal features from DeepFEIVR and a linear combination of causal features from DeepFEIVR-CA, both of which are orthogonal to all linear combinations in previous $i - 1$ canonical correlations. In Figure 7, we provide a canonical correlation plot comparing causal features extracted by DeepFEIVR and DeepFEIVR-CA. It is clear that the top few components are highly

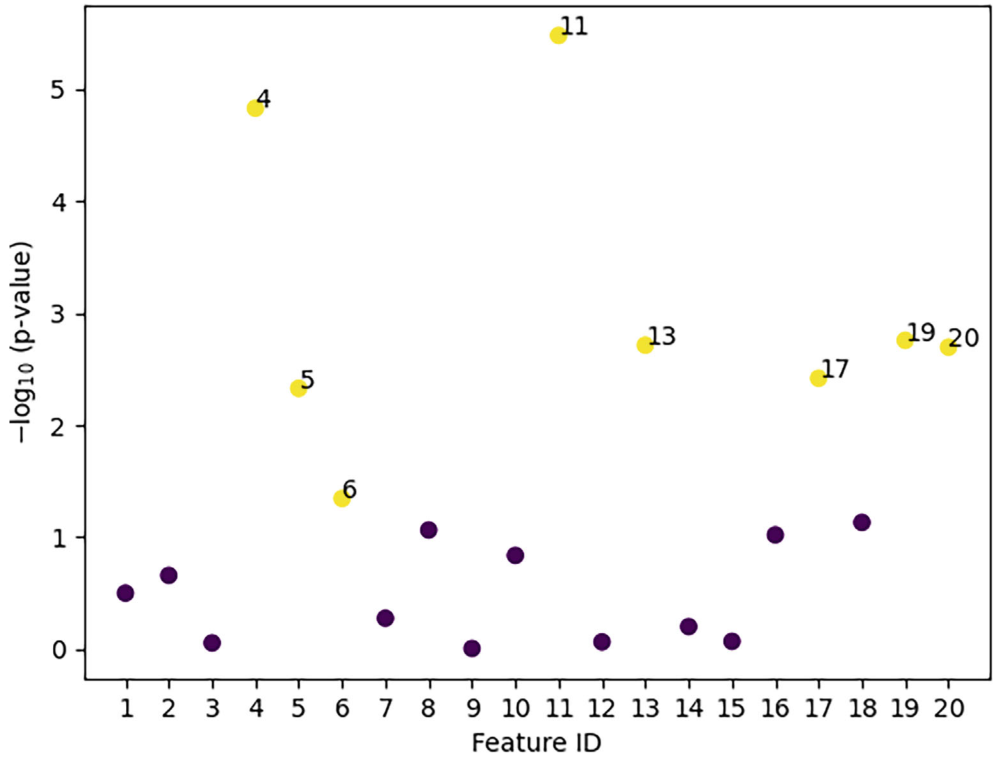

FIGURE 6 Negative $\log_{10}$ (p-values) of individual tests on each of the 20 extracted causal features for its association with AD. The extracted causal features with p-values smaller than 0.05 are labeled with their IDs.

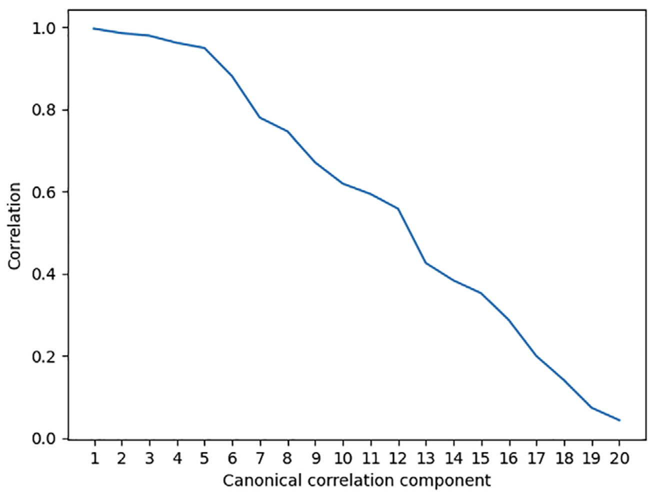

FIGURE 7 The 20 canonical correlations between DeepFEIVR- and DeepFEIVR-CA-extracted features.

related: for example, the first 5 correlations are all larger than 0.9; but on the other hand, some canonical correlations are moderate or small. For these reasons, we will skip further discussions on the results of DeepFEIVR-CA.

4.3 | Extracted features and their interpretation

4.3.1 | Extracted causal features and ROIs

In this part, to facilitate their interpretation, we show how the causal features extracted by DeepFEIVR are related to the brain regions of interest (ROIs). Here we also compare the noncausal features extracted by the direct CNN model and the

TABLE 2 The top 5 most significant ROIs for 3 noncausal (by the direct CNN) or causal PCs (by DeepFEIVR) after PCA.

	Direct CNN	DeepFEIVR
1st	HIPPL	HIPPL
2nd	HIPPR	HESCHLR
3rd	HESCHLR	HIPPR
4th	HESCHLL	TEMPMIDL
5th	ROLANDOPL	TEMPINFL

causal features extracted by DeepFEIVR (see Figure 3 for model differences). In order to compare 20 features together, we extract some top principal components (PCs) from principal component analysis (PCA) of the causal or noncausal features. The number of PCs are determined by the explained proportion of the total variance larger than 10% as we expect each PC to be representative. For PCA models applied to the causal or noncausal features, 3 PCs satisfy this requirement and their cumulaive explained proportions of the total variance are both larger than 70%. Then for each ROI and each PC, we fit a simple linear model, regressing each extracted PC on each ROI and the p-value of the slope coefficient indicates the relatedness between the PC and the ROI. A smaller p-value of the slope coefficient indicates that the extracted PC is more strongly associated with the ROI. In Table 2, we list 5 most significant ROIs associated with the PCs of the features extracted from the direct CNN (noncausal) or DeepFEIVR (causal). The p-value of each ROI is determined as the smallest of its 3 p-values associated with the 3 PCs.

Table 2 lists the top 5 ROIs selected by the two methods: HIPPL (left hippocampus), HIPPR (right hippocampus), HESCHLR (right Heschl's gyrus), HESCHLL (left Heschl's gyrus), ROLANDOPL (left Rolandic operculum), TEMPMIDL (left middle temporal gyrus) and TEMPINFL (left inferior temporal gyrus). Among the above selected ROIs, the hippocampus has long been known to be related to AD in the literature.[37] The left Heschl's gyrus is shown to be helpful in distinguishing AD and healthy individuals.[38] Expression of γ-aminobutyric acid is affected in the middle temporal gyrus of AD individuals.[39] Patients with AD or MCI show a loss of synapses in the inferior temporal gyrus.[40]

Although the top ROIs associated with the causal or noncausal features largely overlap, there are some differences between the three PCs of the causal and noncausal features. Canonical correlations between the two sets of the 3 PCs from the direct CNN and DeepFEIVR are only 0.57, 0.49, and 0.02, respectively.

In Figure 8, we show the plots of $-\log_{10}$ (p-values) between each of 116 ROIs and each of the 3 causal PCs.

Table 2 only considers the brain regions globally associated with PCs. Next, we investigate the ROIs related to each individual feature extracted. In Figure 9, we show the p-values for the associations between each of the 116 ROIs and the 4th or the 13th feature; recall that the 4th and 13th feature are both highly associated with AD based on the IGAP AD GWAS summary statistics. For the 4th feature, the p-values of HIPPL (left hippocampus) and HESCHLR (right Heschl's gyrus) are < 0.05, while for the 13th feature, HIPPL, HIPPR (right hippocampus) and HESCHLR all yield a p-value $< 1 \times 10^{-6}$. However, some important features may not be significantly associated with any ROIs by capturing information beyond the ROIs.

For comparison, we also treat the 166 ROIs as exposures for the outcome AD in multi-variable 2SLS, yielding a p-value of nearly 0 from the global association test. In uni-variable 2SLS on each ROI and AD, HIPPL, HIPPR and HESCHLR are the most significant ROIs, which confirms our earlier findings. Nevertheless, the ROIs contain only information about regional gray matter volumes, but DeepFEIVR can detect other brain features such as multiple ROIs and their interactions or those related to white matter.

4.3.2 | Activation maps of extracted causal features

In this section, we use gradCAM[32] to show how significant features extracted by DeepFEIVR relate to the brain regions. For each significant feature extracted by DeepFEIVR, we can apply gradCAM to this feature and identify the region with the largest absolute activation, which is supposed to contribute most to this feature. Figure 10 gives an example for the activation maps for the 13th feature on the outputs of the second convolutional network block. The absolute activation of each plot is mapped to the range [0, 1]. The active regions appear to be close to the (left and right) hippocampus and other ROIs highlighted in Section 4.3.1.

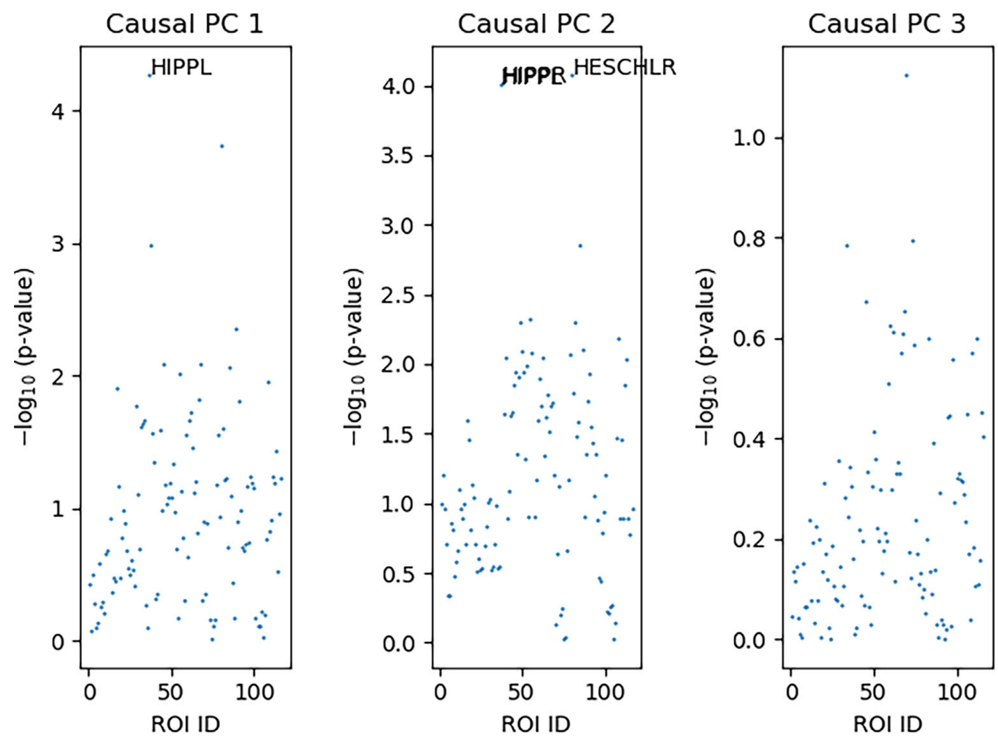

FIGURE 8 Negative $\log_{10}$ (p-values) for associations between each of 116 ROIs and each of the 3 PCs of the extracted causal features.

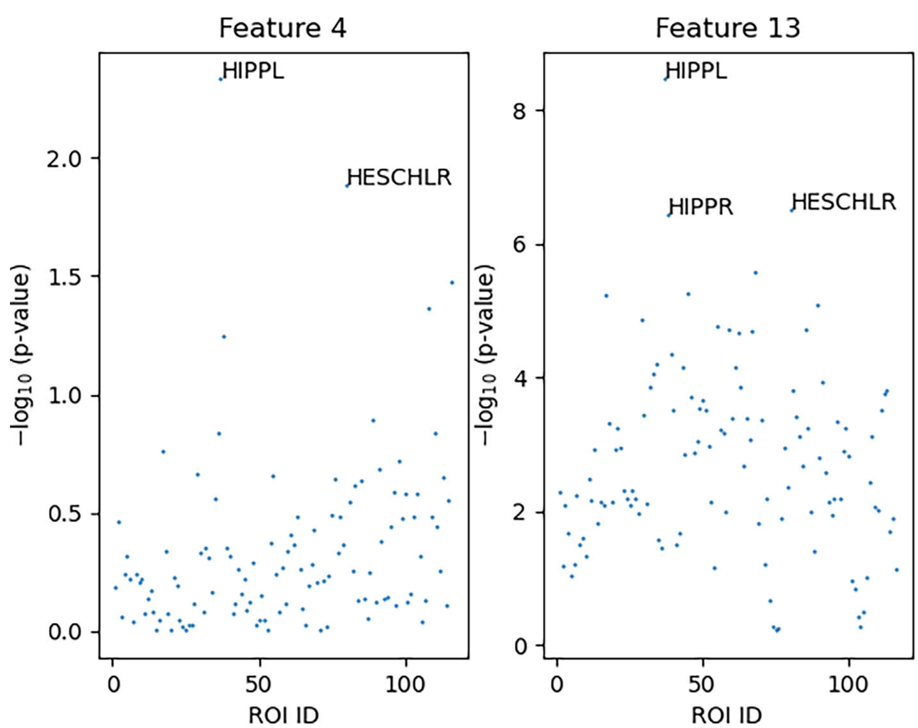

FIGURE 9 Negative $\log_{10}$ (p-values) of associations testing between each of 116 ROIs and the 4th or the 13th extracted causal features.

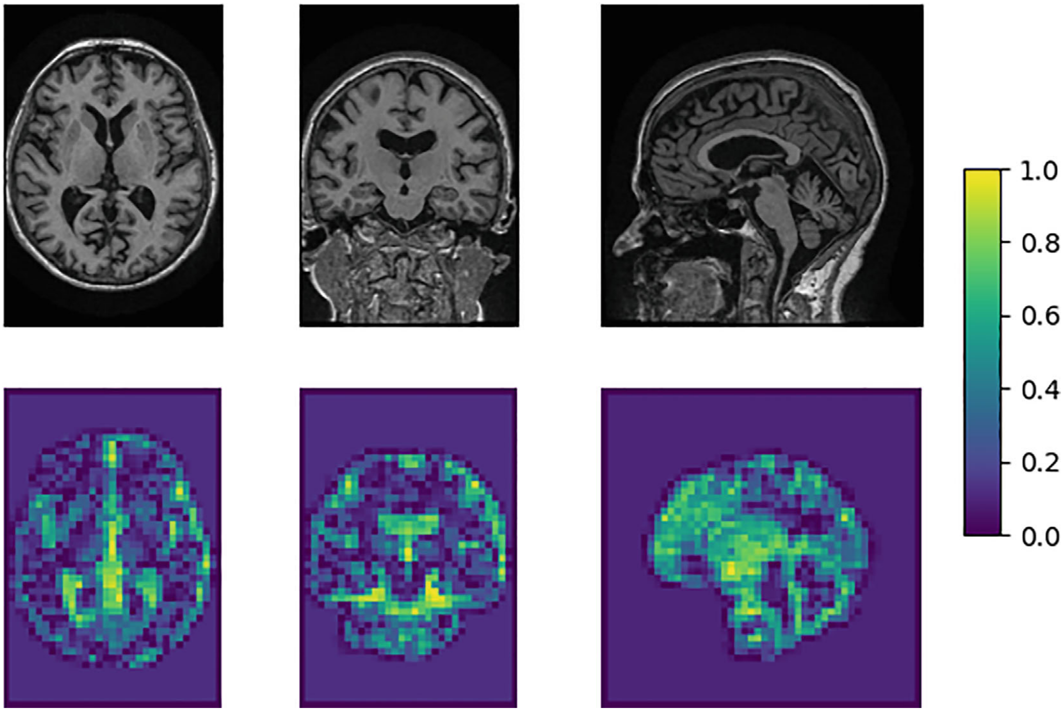

FIGURE 10 GradCAM activation maps for the 13th causal feature.

TABLE 3 The top 10 most significant IDPs.

IDP (abbr.)	IDP
Vol-P-occ	Volume of Pole-occipital (left) by white surface parcellation (Destrieux)
Vol-Somo	Volume of S-orbital-med-olfact (left) by white surface parcellation (Destrieux)
Area-lo-DK	Area of lateral occipital (left) by white surface parcellation (Desikan-Killiany)
Area-V2	Area of V2 (left) by white surface parcellation (BA_exvivo)
Area-lo-DKT	Area of lateral occipital (left) by white surface parcellation (DKT)
Area-P-occ	Area of Pole-occipital (left) by white surface parcellation (Destrieux)
Area-Soml	Area of S-oc-middle+Lunatus (left) by white surface parcellation (Destrieux)
MD-r-ic	Mean diffusivity in retrolenticular part of internal capsule (right) on fractional anisotropy (FA) skeleton
ML1-sup-fo	Mean L1 in superior fronto-occipital fasciculus (left) on FA skeleton
ML3-r-ic	Mean L3 in retrolenticular part of internal capsule (right) on FA skeleton

4.3.3 | Extracted causal features and IDPs

IDPs are some numerical features generated from various types of MRI scans using existing methods. BIG40 publishes the GWAS summary statistics of 3935 IDPs. For each IDP, we conduct a global Wald test for the 20 features extracted by our method. Among 3935 IDPs, 417 IDPs are marginally significant at a p-value threshold 0.05 and 114 IDPs are significant at a p-value threshold 0.01. In Table 3, we list the 10 IDPs with the smallest p-values. Among the 10 selected IDPs, some have been shown to be related to AD in previous literature. Patients with AD show lower fractional anisotropy (FA) and higher mean diffusivity (MD) in internal capsule and fronto-occipital fasciculus.[41] Cortical thickness is lower in lateral occipital cortex of patients with AD.[42]

In addition to the global Wald testing, we conduct individual tests on each of these 10 selected IDPs. We provide a heat map of $-\log_{10}(p\text{-values})$ in Figure 11. We can check the relationships between the extracted individual features and the selected IDPs in Figure 11.

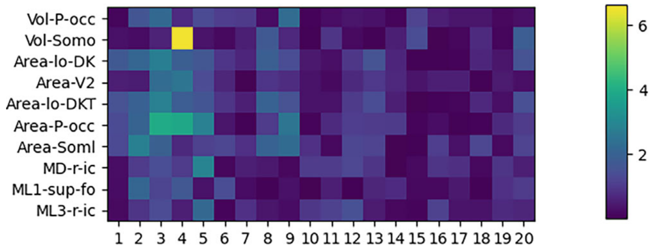

FIGURE 11 The heat map of $-\log_{10}$ (p-values) between 20 features and 10 IDPs.

5 DISCUSSION

In this paper, a new deep learning method called DeepFEIVR is proposed to extract causal features from 3D images. A key feature is its assumption of a linear relationship between the IVs and the extracted (nonlinear image) features/exposure, thus a linear relation between the IVs and the outcome. Some distinct properties of DeepFEIVR are its applicability to high-dimensional data (eg, 3D neuroimages considered here) for causal feature extraction and to GWAS summary data in subsequent extracted feature-outcome association testing. Instead of predicting images directly, we use the extracted features as a linear function of IVs to do prediction during the model training process, ensuring the extracted features are predictive for the outcome while maintaining a causal interpretation. By association testing using the IGAP AD GWAS summary data, we illustrate that some causal features extracted by DeepFEIVR from the ADNI dataset are significantly associated with AD. By gradCAM, we can identify the brain regions contributing most to the extracted features for each input image; and by using brain ROIs and IDPs, we show how the extracted features may be related to some brain regions or IDPs, thus facilitating the interpretation of the extracted features, though this remains a challenging and largely unsolved problem as for any black-box machine learning tools like CNNs. The proposed method DeepFEIVR is also applicable to other large and complex biological data, a part of our on-going work. In the future, we may consider some more general and realistic situations with the presence of invalid IVs, which for example can affect the outcome directly (and thus violating the three valid IV assumptions).

We have extended DeepFEIVR to DeepFEIVR-CA for covariate adjustment. In the presence of covariates, adjusting them can improve the estimation efficiency, though it may not be necessary to do so because they may be treated as hidden confounders. In our real data analysis with the ADNI data, we have considered and adjusted for three covariates, the baseline age, gender and handedness, as often used in previous GWAS analyses of the ADNI data.[7] We have shown that the features extracted from DeepFEIVR-CA are also associated with AD, though less significantly than those from DeepFEIVR, and that the two sets of the features are related. The less significant AD association of the DeepFEIVR-CA-extracted features could be due to that some covariates are involved in the brain-AD causal pathway; for example, it is possible to have gender/handedness ⇒ brain features ⇒ AD; if so, adjusting such a covariate may dilute the effects of some related brain features, and thus increasing the chance of missing these brain features in feature extraction. Another issue is that, since GWAS summary data usually do not provide any information on covariates, it is not possible to conduct hypothesis testing without some additional assumptions. In our real data example, we assume that the IVs are (nearly) uncorrelated with all the covariates, thus ignoring the covariates for hypothesis testing with GWAS summary data would not be a problem. More studies are needed.

Existing neuroimaging GWAS are mostly based on manually extracted imaging features as endophenotypes, for example, some regions of interest (ROIs) based on a brain atlas. However, due to limited knowledge, it is still debatable about how to define ROIs or even brain atlases; furthermore, these ROIs may or may not be most relevant to the given GWAS trait, that is, AD here. For example, there are at least 66 existing atlases for the whole brain structural MRI (sMRI) data;[12] which one is best to use? Given the tremendous successes of CNNs, especially for automatic feature extraction in image analysis, it is natural to apply CNNs to extract features from neuroimaging data, and use these features as the traits (ie, endophenotypes) to be associated with SNPs.[17] These extracted features can go beyond any given ROIs. Here we move one step forward: we'd like CNN-extract image features to be more likely to be causal to the outcome (ie, AD) by taking advantage of IV regression. In addition, given relatively small sample sizes of existing individual-level neuroimaging genetics data (such as the ADNI data), it would be more powerful and thus useful to test extracted features with large-scale GWAS summary data (such as the IGAP AD GWAS summary data), as demonstrated in our real data analyses. In spite of these potential advantages, there are some limitations with the proposed approach. First, it may be time-consuming to develop a suitable CNN architecture, including determining many tuning parameters, for a given problem or dataset. In the current setting, it helps when we can use predictive performance for the outcome (ie, AD status) with a validation dataset to do so. Although there is a literature on neural architecture search (NAS) aiming to automate this searching process, it is still quite computing-intensive.[43] Second, deep learning models are data hungry, often requiring large amounts of data. The sample size of the ADNI, around 800, is still small. Other techniques, such as transfer learning,[44] data augmentation[17] and self-supervised learning can be explored and incorporated in the future for better performance. Third, perhaps most importantly, it is still quite challenging to interpret CNN-extracted features. These are topics warranting future investigations.

ACKNOWLEDGEMENTS

We thank the editors and reviewers for many helpful comments and suggestions. Dataset collection (for the part of the ADNI dataset) in this paper was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). The ADNI is funded by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), as well as receives contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. ADNI clinical sites in Canada are funded by the Canadian Institutes of Health Research. The Foundation for the National Institutes of Health (https://www.fnih.org) is facilitating contributions from private sources. ADNI is organized by the Alzheimer's Therapeutic Research Institute at the University of Southern California and the grantee organization of this study is the Northern California Institute for Research and Education. The Laboratory for Neuro Imaging at the University of Southern California is responsible for distributing the ADNI database.

DATA AVAILABILITY STATEMENT

The ADNI dataset can be accessed by approved users on https://adni.loni.usc.edu. The IGAP summary statistics data are available to approved applicants on https://www.niagads.org/datasets/ng00036. The BIG40 datasets can be downloaded from https://open.win.ox.ac.uk/ukbiobank/big40/. The UK Biobank individual-level GWAS genotype data is available to approved users at https://www.ukbiobank.ac.uk. The code implementing our proposed method is publicly available at https://github.com/yystat01/DeepFEIVRv1.

ENDNOTE

*https://adni.loni.usc.edu/data-samples/data-types/mri/.

ORCID

Xiaotong Shen https://orcid.org/0000-0003-1300-1451
Wei Pan https://orcid.org/0000-0002-1159-0582

REFERENCES

1. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2022;18(4):700-789. doi:10.1002/alz.12638
2. Apostolova LG. Alzheimer disease. *Continuum: Lifelong Learn Neurol*. 2016;22(2 Dementia):419.
3. Beltran JF, Wahba BM, Hose N, Shasha D, Kline RP. Alzheimer's Disease Neuroimaging Initiative tF. Inexpensive, non-invasive biomarkers predict Alzheimer transition using machine learning analysis of the Alzheimer's disease neuroimaging (ADNI) database. *PLOS One*. 2020;15(7):1-26. doi:10.1371/journal.pone.0235663
4. Shen L, Thompson P, Potkin S, et al. Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav*. 2014;8:183-207.
5. Hong M, Reynolds C, Feldman A, et al. Genome-wide and gene-based association implicates FRMD6 in Alzheimer disease. *Hum Mutat*. 2012;33:521-529.
6. Sherva R, Tripodis Y, Bennett D, et al. Genome-wide association study of the rate of cognitive decline in Alzheimer's disease. *Alzheimers Dement*. 2014;10:45-52.
7. Shen L, Kim S, Risacher S, et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. *Neuroimage*. 2010;53:1051-1063.

8. Zhang Y, Xu Z, Shen X, Pan W. Alzheimer's Disease Neuroimaging Initiative f. testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *Neuroimage*. 2014;96:309-325.
9. Zhao B, Luo T, Li T, et al. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat Genet*. 2019;51:1637-1644.
10. Zhao B, Li T, Yang Y, et al. Common genetic variation influencing human white matter microstructure. *Science*. 2021;372:eabf3736.
11. Zhu H, Li T, Zhao B. Statistical learning methods for neuroimaging data analysis with applications. 2022.
12. Dickie DA, Shenkin SD, Anblagan D, et al. Whole brain magnetic resonance image atlases: a systematic review of existing atlases and caveats for use in population imaging. *Front Neuroinform*. 2017;11:1.
13. Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. *J Am Stat Assoc*. 2013;108(502):540-552. doi:10.1080/01621459.2013.776499
14. Miranda MF, Zhu H, Ibrahim JG. TPRM: Tensor partition regression models with applications in imaging biomarker detection. 2015.
15. Oh K, Chung YC, Kim KW, Kim WS, Oh IS. Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning. *Sci Rep*. 2019;9(1):18150. doi:10.1038/s41598-019-54548-6
16. Patel K, Xie Z, Yuan H, et al. New phenotype discovery method by unsupervised deep representation learning empowers genetic association studies of brain imaging. *medRxiv*. 2022. doi:10.1101/2022.12.10.22283302
17. Chakraborty D, Zhuang Z, Xue H, Fiecas MB, Shen X, Pan W. Deep learning-based feature extraction with MRI data in neuroimaging genetics for Alzheimer's disease. *Genes*. 2023;14(3):626. doi:10.3390/genes14030626
18. Klungel O, Uddin MJ, Boer dA, et al. Instrumental variable analysis in epidemiologic studies: an overview of the estimation methods. *Pharm Anal Acta*. 2015;6(353):2.
19. Mo C, Ye Z, Ke H, et al. A new Mendelian randomization method to estimate causal effects of multivariable brain imaging exposures. *Pac Symp Biocomput*. 2022;27:73-84.
20. Taschler B, Smith SM, Nichols TE. Causal inference on neuroimaging data with Mendelian randomisation. *Neuroimage*. 2022;258:119385. doi:10.1016/j.neuroimage.2022.119385
21. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*. 2013;37(7):658-665. doi:10.1002/gepi.21758
22. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol*. 2015;181(4):251-260. doi:10.1093/aje/kwu283
23. Hall P, Horowitz JL. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*. 2005;33(6):2904-2929. doi:10.1214/009053605000000714
24. Newey WK, Powell JL. *Instrumental variable estimation of nonparametric models*. 2003;71(5):1565-1578.
25. Dai B, Li C, Xue H, Pan W, Shen X. Inference of nonlinear causal effects with GWAS summary data, https://arxiv.org/abs/2209.08889 2022.
26. He R, Liu M, Lin Z, Zhuang Z, Shen X, Pan W. DeLIVR: a deep learning approach to IV regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics*. 2023. doi:10.1093/biostatistics/kxac051
27. Xu L, Chen Y, Srinivasan S, Freitas dN, Doucet A, Gretton A. Learning Deep Features in Instrumental Variable Regression. 2020.
28. Hartford J, Lewis G, Leyton-Brown K, Taddy M. Deep IV: a flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research.* 2017;70:1414-1423.
29. Bennett A, Kallus N, Schnabel T. Deep generalized method of moments for instrumental variable analysis. *Adv Neural Informat Process Syst*. 2019;32.
30. Knutson KA, Deng Y, Pan W. Implicating causal brain imaging endophenotypes in Alzheimer's disease using multivariable IWAS and GWAS summary data. *Neuroimage*. 2020;223:117347. doi:10.1016/j.neuroimage.2020.117347
31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vision*. 2019;128(2):336-359. doi:10.1007/s11263-019-01228-7
32. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp*. 2002;17(3):143-155. doi:10.1002/hbm.10062
33. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*. 2001;20(1):45-57.
34. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*. 2002;15(1):273-289. doi:10.1006/nimg.2001.0978
35. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. 2013;45(12):1452-1458. doi:10.1038/ng.2802
36. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):1-10. doi:10.1371/journal.pmed.1001779
37. Setti SE, Hunsberger HC, Reed MN. Alterations in hippocampal activity and Alzheimer's disease. *Transl Issues Psychol Sci*. 2017;3(4):348-356. doi:10.1037/tps0000124
38. Hänggi J, Streffer J, Jäncke L, Hock C. Volumes of lateral temporal and parietal structures distinguish between healthy aging, mild cognitive impairment, and Alzheimer's disease. *J Alzheimers Dis*. 2011;26(4):719-734. doi:10.3233/JAD-2011-101260
39. Govindpani K, Turner C, Waldvogel HJ, Faull RLM, Kwakowsky A. Impaired expression of GABA Signaling components in the Alzheimer's disease middle temporal gyrus. *Int J Mol Sci*. 2020;21(22):8704. doi:10.3390/ijms21228704
40. Scheff SW, Price DA, Schmitt FA, Scheff MA, Mufson EJ. Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's disease. *J Alzheimers Dis*. 2011;24(3):547-557. doi:10.3233/JAD-2011-101782

41. Mayo CD, Garcia-Barrera MA, Mazerolle EL, Ritchie LJ, Fisk JD, Gawryluk JR. Relationship between DTI metrics and cognitive function in Alzheimer's disease. *Front Aging Neurosci*. 2018;10:436. doi:10.3389/fnagi.2018.00436
42. Yang H, Xu H, Li Q, et al. Study of brain morphology change in Alzheimer's disease and amnestic mild cognitive impairment compared with normal controls. *Gen Psychiatr*. 2019;32(2):e100005. doi:10.1136/gpsych-2018-100005
43. Liu H, Simonyan K, Yang Y. DARTS: Differentiable Architecture Search. 2019.
44. Dhinagar NJ, Thomopoulos SI, Rajagopalan P, et al. Evaluation of transfer learning methods for detecting Alzheimer's disease with brain MRI. *bioRxiv*. 2022. doi:10.1101/2022.08.23.505030

How to cite this article: Yao Y, Charkraborty D, Zhang L, Shen X, Alzheimer's Disease Neuroimaging Initiative, Pan W. Deep causal feature extraction and inference with neuroimaging genetic data. *Statistics in Medicine*. 2023;42(20):3665-3684. doi: 10.1002/sim.9824

APPENDIX

Details of CNNs implemented on simulation and ADNI data

Figure A1 shows the model architecture of DeepFEIVR applied to the simulated dataset, which involves 2 convolutional neural network (CNN) layers, a global averaging pooling (GAP) layer as well as 2 fully connected layers. Finally, we project the q features (the output of a leaky ReLU layer) into the space of IVs to extract q causal features and a final fully connected layer combines q extracted features for prediction. In Figure A1, CONV, MP, BN, GAP and FC represent a convolutional layer, a max pooling layer, a batch normalization layer, a global averaging pooling layer as well as a fully connected layer. CONV(N@a×a) represents a convolutional layer with N filters and convolution size (a, a) and MP(a×a) represents a max pooling layer with pooling size (a, a). FC(N) represents a fully connected layer with N neurons.

The left panel of Figure A2 shows a CNN module to estimate f using a direct CNN model as well as DeepFEIVR, which involves four 3D convolutional neural network layers and three fully connected layers. After each CNN layer, a 3D max pooling (MP) layer and a batch normalization (BN) layer are added. We also add a global averaging pooling (GAP) layer after the last CNN layer and two dropout layers between three FC layers. In the right panel of Figure A2, we compare a direct CNN model and DeepFEIVR. In the direct CNN model, a linear regression model is directly applied to the output of the CNN module while DeepFEIVR involves a projection layer.

The architecture of DeepFEIVR-CA for ADNI data is in Figure A3. We project the features extracted by the CNN into the column space of the IVs and covariates. Finally, we use a fully connected layer applied to a concatenated vector of the projected features and covariates for prediction.

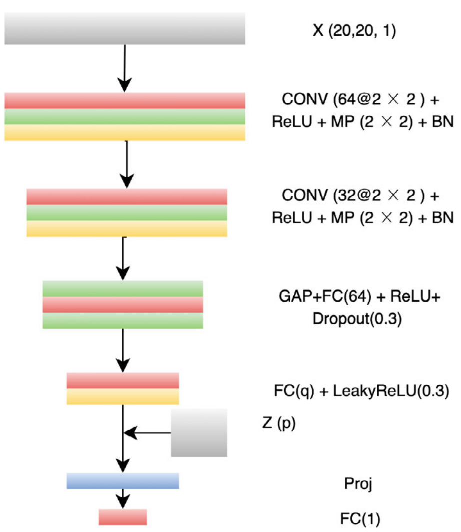

FIGURE A1 The model architecture of DeepFEIVR applied to the simulation dataset.

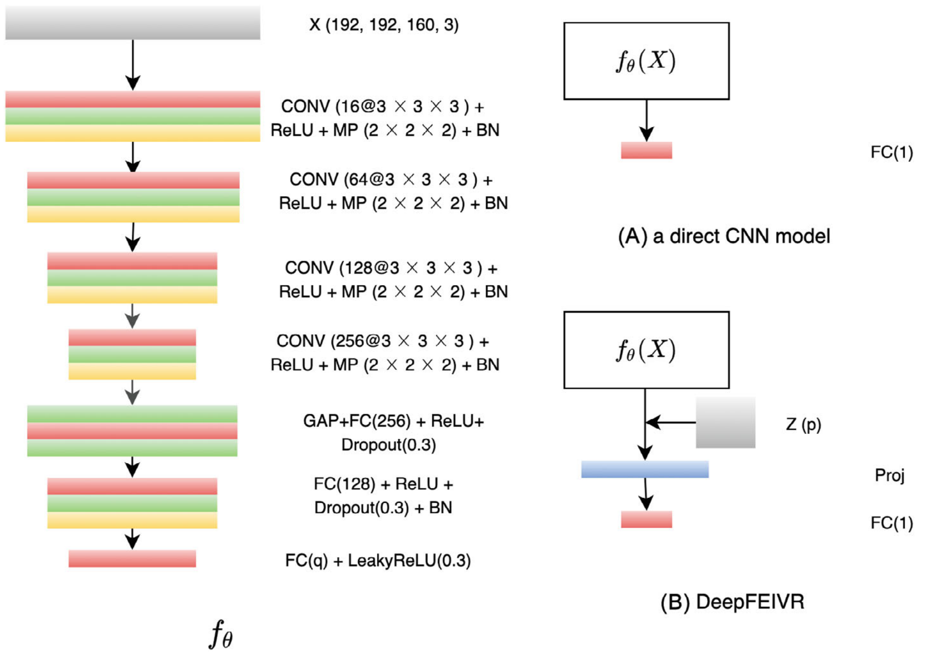

FIGURE A2 Models used for the ADNI dataset. Left: the model architecture f_θ. Top-Right: the direct CNN model in which a linear regression model follows f_θ. Bottom-Right: DeepFEIVR in which the extracted features from f_θ are projected onto the column space of the IVs, then a linear regression model is applied.

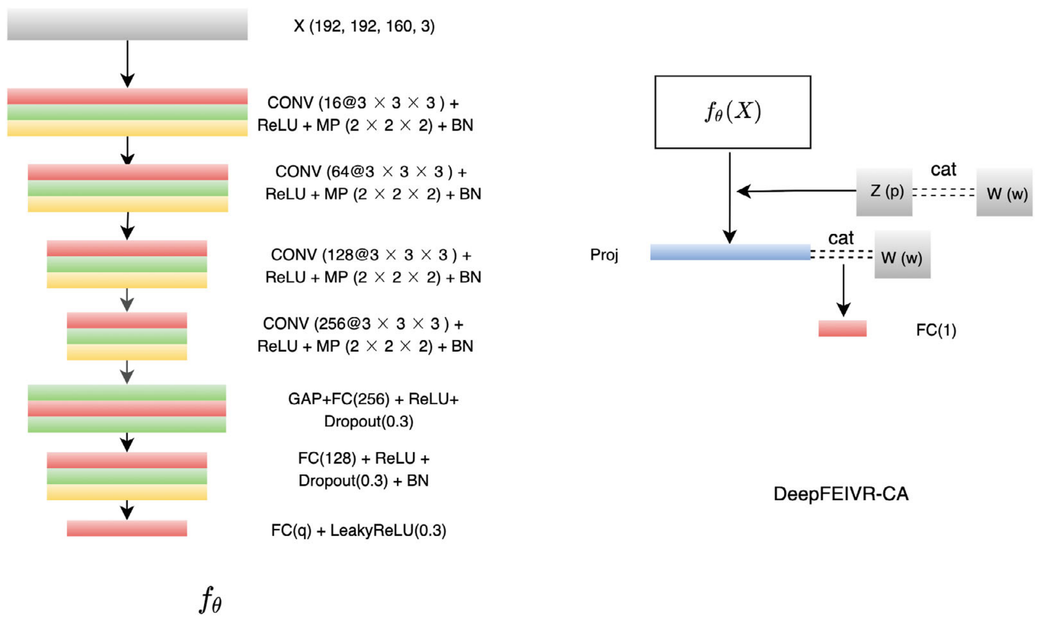

FIGURE A3 The model architecture of DeepFEIVR-CA used in the ADNI dataset. Left: the model architecture f_θ. Right: DeepFEIVR-CA.